IDENTIFICATION OF ELECTORAL IRREGULARITIES BY STATISTICAL METHODS

Bochsler, Daniel; Medzihorsky, Juraj; Schürmann, Carsten; Stark, Philip

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Strasbourg, 19 March 2018 **CDL-AD(2018)009**

Opinion n° 874 / 2017 Engl. only

EUROPEAN COMMISSION FOR DEMOCRACY THROUGH LAW

(VENICE COMMISSION)

REPORT

ON THE

IDENTIFICATION OF ELECTORAL IRREGULARITIES BY STATISTICAL METHODS

Taken note of by the Council for Democratic Elections at its 61st meeting (Venice, 15 March 2018) and the Venice Commission at its 114th Plenary Session (Venice, 16-17 March 2018)

on the basis of comments by:

Mr Daniel BOCHSLER (Expert, Denmark)
Mr Juraj MEDZIHORSKY (Expert, Sweden)
Mr Carsten SCHÜRMANN (Expert, Denmark)
Mr Philip STARK (Expert, United States of America)

Contents

I. Introduction

1. At its meeting of December 2016, the Council for Democratic decided to prepare a report on the identification of electoral irregularities by statistical methods It entrusted four experts, Messrs Daniel Bochsler, Juraj Medzihorsky, Carsten Schürmann and Philip B. Stark, to provide their contributions.

2. At the 59th meeting of the Council for Democratic Elections (June 2017), Mr Schürmann developed the main lines of the envisaged report, which were discussed by the Council. The members of the Council underlined the interest of electoral observers to be acquainted with the issue. The draft report was discussed by the Council at its 61st meeting (March 2018)

3. The Council for Democratic Elections took note of the present report at its 61st meeting (Venice, 15 March 2018), as well as the Venice Commission at its 114th Plenary Session (Venice, 16-17 March 2018).

II. Executive Summary

4. The report summarises exemplar statistical tests of numerical election results, but also shows that such tests must be combined with other types of observations, informed by country-specific expertise—and are still fallible then. While there is a broad range of statistical tests for election irregularities, the tests cannot be relied upon exclusively: they are subject to unknown rates of false positives and false negatives. The difficulty quantifying error rates is in part because statistical election forensic tests rely on probability models for the results of clean elections, and those models are largely heuristic, because there is no “physics” of political preference and voting behaviour against which to check numerical results.

5. Democratic elections are more than counting votes. They are supposed to be free, fair, and inclusive competitions that allow voters to “vote their conscience.” Fairness depends on many factors. For instance, an election where the authorities only allow candidates who are supportive of the regime to run for office will still be a mockery of an election, even if voters pick their personal favourites from the list without fear of reprisal, votes are counted transparently and accurately, and the results are reported accurately.

6. To understand the variety of electoral malpractice, we first need to understand what is necessary for elections to be free and fair. The components of “electoral integrity” are addressed in a number of documents by the European Commission for Democracy Through Law (see for example the Venice Commission Code of Good Practice in Electoral Matters (CDL-AD(2002)023rev2) and its Report on Figure Based Management of Possible Election Fraud (CDL-AD(2010)043). The International Covenant on Civil and Political Rights, in Article 25, implies that genuine elections are based on the right to stand for public office and contest elections; universal suffrage; a catalogue of civil rights and freedoms, in particular those necessary to conduct an electoral campaign (freedom of information, assembly and association); equal suffrage, i.e. votes of all voters contribute equally to the result; the use of a secret ballot process; and the prevention of corruption. The European electoral heritage, as enshrined the first Additional Protocol to the European Convention on Human Rights and the Code of Good Practice in Electoral Matters drafted by the Venice Commission, includes universal, equal, free, secret and direct suffrage, elections at a regular interval, as well as the conditions for implementing these principles - respect for fundamental rights, stability of electoral, organisation of election by an impartial body, observation of elections, an effective system of appeal).

7. Statistical methods for detecting election irregularities from the numerical results cannot hope to find problems with many of these factors. They are best suited to detect issues with the

counts that might result from fraud, manipulation, or error. They generally work by assuming that "fair" and accurate election results would follow a particular hypothesised statistical model, then looking for evidence that the results in question do not follow that model. That is, they assume that the results are correct and that they were generated by a known, understood statistical process, absent evidence to the contrary.

8. The most rigorous and reliable methods for checking whether electoral outcomes are correct (i.e., whether the reported winners actually won) frame the problem in the opposite way: the electoral outcome is assumed to be incorrect (in an unknown way), absent convincing evidence that the reported winners really won. These "risk-limiting audit" methods can generate affirmative statistical evidence that results are correct, guaranteeing that if the result is incorrect, the chance the result would escape correction is at most some pre-specified value α (and the chance the audit will correct the result is at least 1 − α), for any desired α between 0 and 100% (Lindeman and Stark, 2012).

9. However, risk-limiting audits results require more than the vote tallies as evidence. They involve manually inspecting a random sample of records from a voter-verifiable paper trail of voters' preferences (ideally, voter-marked paper ballots), and require evidence that the paper trail is complete and intact (Stark and Wagner, 2012). To conduct a risk-limiting audit, the auditors must be able to draw a random sample of ballots and must have physical access to the selected ballots. The probability calculations involve worst-case assumptions about election results--they provide guarantees whether errors are random or are introduced by a malicious opponent--and are based the known (by fiat) probability that the audit will select any particular set of paper records for inspection.

10. In contrast, forensic methods that rely only on the reported numerical results cannot check whether an electoral outcome is correct, nor offer any statistical guarantees. However, they can detect some ways in which the numerical results might be tainted by fraud or error, and they can be used even when the investigators do not have access to accurate underlying voter-verified records--because the investigators lack legal standing (e.g., they are members of the public, employees of NGOs, or foreign election observers), because the paper records do not exist (e.g., when voter preferences are captured using direct-recording electronic voting machines that do not generate a voter-verifiable paper trail), or because the paper records are not reliable (e.g., the chain of custody of the ballots might have been compromised, ballot boxes might have been "stuffed," or ballots might have been lost, added, substituted, or altered).

III. General remarks

11. Over the last few decades, most nations have adopted some form of elections. (Kofi Annan Report, 2012) This has brought about increasing concern that many of these elections are not in conformity with international standards. In the 1960s and 1970s, only one in six elections worldwide was criticized for large-scale electoral fraud or manipulation. Since the 2000s, this rate has almost doubled to about three in ten elections. Controversies about the legitimacy and integrity of elections can lead to turmoil, violence, and even civil war (Daecker, 2012, Laakso, 2007, Brancati and Snyder, 2013). This has led a growing number of researchers and policy-makers to develop methods to assess the integrity of elections. In particular, the Venice adopted in 2010 the Report on Figure Based Management of Possible Election Fraud (CDL-AD(2010)043), which addresses the ways of detecting fraud throughout the various stages of the electoral process.

12. Statistical methods are one class of checks.

13. The technical and mathematical nature of statistical tests raise the expectation that they provide an objective standard for fair elections, checking results even where electoral

observation missions are unable to provide “boots on the ground.” Observers can never be present throughout the electoral process in all relevant places. They might even be prevented from accessing some polling stations, or become involved in the turmoil of political accusations. In contrast, as long as results are available, statistical tests might have something useful to say.

14. Unfortunately, statistical tests have subjective elements of their own, and are quite fallible: the risk of misjudging an election result by statistical tests is high. Statistical tests can falsely identify legitimate election results as irregular, and can fail to flag a fraudulent result as such, especially if the agent of fraud is careful and knows the general types of tests that will be used. Therefore, statistical test results must be interpreted with great circumspection.

15. The following examples give the flavour of some statistical tests.

Example 1:
16. There is nothing suspicious about a polling station reporting 70.01% of the votes cast for the ruling party. However, if many polling stations in the same country report round percentage numbers, e.g. 70.0% or 70.1% for the ruling party, and far fewer vote shares below 70% or above 71%, this might suggest fraud. Political authorities might have set a 70% vote share target for the ruling candidate or party (or another round percentage value), and the local authorities may have delivered the target by altering ballots, stuffing ballot boxes, or altering the count (Kobak et al., 2016a).

Example 2:
17. It has been repeatedly found that when individuals make up numbers they tend to pick some digits too frequently and others too rarely. Applied in the field of elections, if results from polling stations feature certain decimals much more frequently than others, this might indicate that these results are not the product of a genuine vote count, but instead were invented or manipulated (Mebane, 2008, Deckert et al., 2011, Medzihorsky, 2015, and many others).

Example 3:
18. In every election, turnout varies across polling stations. If turnout is strongly associated with vote share in favour of one party, this might indicate that the electoral result in some areas (e.g., where manipulation is easier) was fraudulent (Kobak et al., 2016b).

19. In each of the three examples, the pattern in the results might signal electoral malpractice. If in some parts of the country–e.g., the periphery–voters are coerced to vote, and to vote for one particular candidate or party, then the affected polling stations will show high turnout and strong support for that candidate or party. The same pattern emerges if election officials in some polling stations engage in “ballot-box stuffing”, adding ballots for their favourite candidate to the ballot boxes. Peculiar patterns of numbers, where some digits are present unnaturally often, might indicate that tallies have been massaged.

20. Statistical methods can identify such anomalies in election results. However, they can neither show definitively whether elections have been subject to illegitimate manipulation, nor can they confirm that elections have been conducted correctly and fairly. Instead, the interpretation of statistical tests, and possible confounding factors, requires sound knowledge of the electoral process in the country under study, and of the electoral geography (Myagkov et al., 2009: 267, Leemann and Bochsler, 2014). Statistical analyses can complement and improve the assessment of elections if they are employed in combination with related information, e.g. from election specialists, knowledge of the electoral geography, and qualitative observation of the electoral process by media, civil society or international observers.

21. This report reviews some types of election manipulations and how they might be detected with new statistical approaches to election forensics. It discusses the limits of statistical forensic

approaches, makes suggestions about their application in practice, and provides recommendations to practitioners.

IV. Electoral Malpractice

22. In this report, we consider any manipulation or error that undermines the integrity of elections as free, fair, and competitive to be an electoral irregularity. There are many ways election results can be manipulated or otherwise erroneous, and thus it is natural that there is a large range of forensic procedures to identify different kinds of electoral irregularities. Among the many statistical procedures to vet electoral results, most address relatively few types of irregularities. The most common tests look for evidence that election results were fabricated or altered by election officials.

23. Patterns in numbers fabricated by humans tend to differ from those that occur "naturally" in genuine counts. Statistical procedures have been developed that are sensitive to those differences (but they cannot perfectly distinguish genuine results from fabricated results).

24. A second family of tests looks at turnout, invalid votes, and uniformity of vote shares across polling stations or political geography. Variation across polling stations is natural, and should not automatically be attributed to malpractice. However, there are statistical procedures that test whether patterns (e.g. high turnout and high uniformity of the vote) would be surprising if the votes were generated by a particular, hypothetical "fair" process.

25. Just as electoral integrity requires an extensive set of steps and conditions, failings of electoral integrity can occur in many ways. Election fraud is most often associated with irregularities on election day, including suppressing or intimidating voters, stuffing ballot boxes with extra ballots, altering the contents of ballot boxes, deliberately miscounting the votes, or misreporting the counts. However, election integrity can also be compromised in other parts of the election process. Schedler (2002), list of possible malpractices, including

- restrictions or de-facto hurdles in the registration of candidates or voters
- restrictions on the right to assembly, on campaigns, or the use of public resources and media for campaigning
- use of coercion or threats to affect citizens' participation in elections
- vote buying
- partisan bias in the electoral rules
- alteration of the ballot or ballot-box stuffing
- rigging the election count or the reporting of the count
- preventing elected officials from taking office
- depriving elected bodies of their decision making power

26. Some instances of fraud or manipulation occur in central election administrations, others are decentralised. While most manipulations are committed by the public authorities, other actors (e.g. service providers, hardware manufacturers, paramilitaries, pro-regime parties, opposition groups, hackers, or foreign governments) also can also have a role in election manipulation.

27. Manipulation can be a major impediment to the integrity of elections, and can "legitimize" a government that does not represent the preferences of the people. A wide range of methods to assess the integrity of elections has been established both for political practice and for academic purposes (Norris et al., 2014). After the end of the Cold War, internationally staffed election observation missions became increasingly common (Hyde, 2011), and their reports have been analysed by academics to produce systematic measures of electoral manipulation (Kelley and Kolev, 2010). Researchers also rely on country experts (Norris, 2015), media reports or fraud allegations in court (Alvarez and Boehmke, 2008), or a combination of all three

(Hyde and Marinov, 2011). However, such methods tend to be sensitive to partisan messaging about election fraud.

28. Recently, a number of researchers have focused on statistical methods to identify anomalies in election results (among others Myagkov et al., 2009, Mebane, 2008, Deckert et al., 2011, Hicken and Mebane, 2015). This report reviews these methods below. The general technique to identify electoral irregularities by statistical means is to select a data source, formulate a hypothesis about how election results would be generated statistically in the absence of fraud, and process the data using what is called a statistical test.

29. The data these tests rely on are of two general kinds. There are *observable* data sources, included vote tallies at polling or constituency levels, data on valid and invalid votes, number of valid votes, voter turnout data, but also national registries, including census, national register of birth and deaths, etc. And there are *non-observable* data sources, which may be hidden from those conducting the analysis, for example, voter registration lists, transmission logs, voter turnout percentages, lists of observers and party agents, access restrictions to polling places, and so on.

30. One can only test hypotheses about observable data sources. Some hypotheses, however, may require mentioning non-observable data-sources. For example, suppose there is a polling centre with three voting stations, and two polling stations show candidate A as winner with 60.1% and 58.7%, respectively, whereas the third shows B as a winner with 55%. This polling centre looks suspicious, assuming that voters are assigned uniformly to polling stations, because we would have expected comparable voting preferences across all three polling stations. If voters are not uniformly distributed among polling stations but alphabetically, it may very well be that several families all voting for B were registered with the same third polling place, providing a benign explanation of the discrepancy. For a statistical analysis, the data source of valid invalid votes for each polling station is observable, but to become a testable hypothesis, non-observable information about on how voters are allocated to polling station must also be considered.

31. The final factor of any election manipulation analysis is knowledge of the adversaries and their capabilities. An adversary could be a ruling party that seeks to stay in power; it may be a presiding officer in a polling station, who would like to alter the result; or it might be a foreign nation state that seeks to disrupt the government or install its own friendly government. If the ruling party is under suspicion for election manipulation, their capabilities generally exceed those of any single presiding officer, so they could in principle do much greater damage to the election outcome. The use of electronic election technologies might enable a single actor (including a foreign power) to completely control election results, absent a durable, tamper-evident, voter-verified paper trail that is used to check the electronic results by a rigorous audit.

V. Statistical Testing

32. Statistical tests have been developed to analyse detailed electoral results, and to identify odd or surprising patterns in these numbers. Irregularities in electoral results might be the result of manipulations in the electoral process or during the vote count. The three following examples represent different types of fraud that can be detected with an appropriate statistical procedure, designed to be sensitive to that particular form of electoral malpractice.

33. Most work has been on tests that focus on the vote counting process to detect patterns that are deemed unlikely to occur in genuine counts, but that might be a by-product of fabricating results. A second family of tests looks for surprising correlations between different aspects of the voting results. The range of fraud this second family of tests might uncover is more comprehensive. For example, they might detect systematic ballot invalidation during the vote

counts in some areas, larger patterns of ballot-box stuffing by election officials, or even local coercion of voters to turn out and vote for one party. Thus, this second family might detect electoral malpractice that occurred on election day, not just in the tally. However, these tests rely on stronger assumptions about the mechanism that generates un-manipulated election results. In turn, that entails a higher risk of false alarms from patterns that arose from natural processes, rather than fraud. This report will discuss these limits below.

34. Because malpractice can alter election outcomes at many points in the electoral process, there is a need for a spectrum of statistical tests sensitive to problems in different aspects of the election. Unfortunately, using a multitude of tests also increases the risk of falsely flagging a fair election as fraudulent, as a result of "statistical multiplicity."

35. Nonetheless, there are some types of electoral malpractice that cannot be detected by statistical tests, and which only can be identified through qualitative information. This includes unfair conditions for election campaigns; irregularities in the registration of candidates or voters; systematic national practices of ballot invalidation or other practices affecting the voting results throughout the territory; and limitations on the exertion of the mandate.

A. Statistical Detail

36. Deciding whether electoral results are incorrect or have been manipulated, altered, or falsified amounts to a (binary) classification problem: label a given election as "clean" or as "tainted." "Tainted" might mean that the electoral outcome is wrong--i.e., that the reported winner(s) did not actually win--or merely that the reported numbers are wrong, depending on context. "Clean" might mean that the reported numbers are accurate, or that the announced winners are the true winners, despite any errors or malfeasance that might have occurred.

37. In every binary classification problem, two kinds of errors can occur, namely, misclassifying an item that belongs in the first class as belonging to the second class, and vice versa. In statistics, binary classification is often formalised as a hypothesis test. A natural framing of the classification problem in election forensics would state the null hypothesis to be that the results are clean, and the alternative hypothesis to be that the results are tainted. A false positive, also known as a Type I error, is to label a clean election "tainted"; a false negative, also known as a Type II error, is to label a tainted election "clean."

38. If there were a reliable statistical model for how election results were generated in the absence of error or manipulation (a "generative model" for fair election results), one could develop methods that have a known maximum chance of misclassifying a "clean" election as "tainted," by posing the classification problem as an hypothesis test as described above. The null hypothesis is that the election is clean; in the alternative, it is tainted. One could then construct tests that have probability at most α of concluding that a clean election is tainted, for any desired α between 0 and 100%. If there were a reliable generative model for election results in the presence of error and/or manipulation, one could determine the chance any particular method would misclassify a tainted result as clean and could seek tests that maximise the power to detect fraud or error of various kinds.

39. Unfortunately, the probabilities that arise in numerical forensic methods are not correct simply by fiat, as the probabilities in risk-limiting audits are. That is because the probabilities do not arise from the auditor selecting a random sample in a controlled, deliberate way, or from any other mechanism known to or controlled by the investigator. Rather, the calculated probabilities derive from assumptions about the probability distribution of clean election results: how the numbers "should" behave absent error, fraud, or manipulation, and how that behaviour does or does not vary with other factors.

40. These statistical assumptions are not grounded in established, fundamental truths about political preferences: there is no reliable "physics" of elections. Rather, the assumptions are more-or-less plausible working hypotheses that lead to predictions that can be compared to data.

41. Evidence that reported results are statistically inconsistent with those assumptions presents a mystery: why does this set of results seem peculiar? The solution to that mystery need not be election irregularity; it could simply be that those more-or-less plausible assumptions are not (approximately) true in that election. For instance, many irregularity-detection methods involve assumptions about the distribution of digits in correct election results. One class of methods assumes that in the absence of irregularity, the terminal digits would be uniformly distributed and independent from locality to locality. Another class of methods assumes that in the absence of irregularity, the leading digits of subtotals should follow Benford's law, according to which the fractional part of the logarithm of election results is uniformly distributed between 0 and 1 and independent from locality to locality.

42. Still other methods make assumptions about political preferences, for instance, that the vote share a candidate receives does not depend systematically on the number of votes cast in the polling place, or that political preferences do not depend systematically on the technology a jurisdiction uses to count votes. Demonstrating that such hypotheses are not consistent with the data does not imply that there was any irregularity (Lindeman, 2015).

43. Conversely, a skilled adversary can always fabricate fraudulent results in such a way that the numbers will pass any given suite of statistical tests. There is evidence that this occurred in a case where the mean of the second digits fits the predictions of Benford's Law essentially perfectly, and the mean of the terminal digits fits a uniform model essentially perfectly (Kalinin & Mebane, 2017). Kalinin and Mebane conclude that this agreement is too good to have occurred naturally, and argue that it is a deliberate signature by fraudsters. More generally, there is no such thing as a statistical test that has high power against all alternatives (Freedman, 2010).

B. Statistical Multiplicity

44. Whenever a suite of statistical tests is applied to the same data, the chance of at least one "false positive" result is generally larger than the chance of a false positive for each individual test considered singly. This is called "multiplicity." The tendency for multiplicity to increase the false positive rate is exacerbated when the tests are selected after examining the data. (For instance, if a forensic investigator notices that the results frequently end in 0 or 5, then decides for that reason to test whether the occurrence of 0 and 5 is surprising, the chance of a false positive increases substantially.)

45. Given enough different tests, it is virtually certain that the data will fail at least one. The implication for election forensics is that if we use a suite of many tests to examine election results, the chance that at least one test will flag the results as suspicious may be large, even if the election is "clean" and each test is individually unlikely to raise a false alarm.

46. As an illustration, suppose we wish to test whether the terminal digit behaves as if it is random, in particular, whether all digits are equally likely to occur and whether there is any connection (dependence) among the final digits in different reporting groups. The statistical question is whether the final digits are "independently and uniformly distributed on {0, 1, ..., 9}." There are a variety of standard statistical tests, but they are sensitive to different kinds of anomalies. For instance, we might test using the mean of the terminal digits (which is expected to be 4.5 if the digits are indeed uniformly distributed), or using the Kolmogorov test of the empirical distribution against the theoretical probability mass function that assigns chance 0.1 to

each possibility; or using a chi-square test for equal frequencies of all digits[1]; or using the multinomial range test to compare the most frequent and least frequent digits; or a test for multimodality; or a test based on the frequency with which 0s and 5s occur; or any number of other tests. If we use enough such tests, and especially if we examine the data before choosing which tests to apply, it can be quite likely that at least one will classify the election as "tainted."

47. Bonferroni's inequality is a conservative way to account for multiplicity in testing: the combined significance level (chance of a false positive) of a collection of tests is at most the sum of their significance levels, even when the tests have arbitrary statistical dependence (provided all the tests considered are included in the sum--Bonferroni's inequality does not protect against false positives if the tests are chosen after looking at the data). The chance of an arbitrary union of events is at most the sum of their separate chances. However, because it gives universal protection, Bonferroni's inequality can be quite conservative in some circumstances.

C. Primer on Testing

48. Digit-based election forensics (DBEF) tries to determine whether an election was fraudulent by looking at the numerical results at the finest level at which vote counts were reported. This is appealing particularly because of its low cost, especially when local vote counts are available online. The methods are designed to classify a set of election returns as fraudulent or fraud free by inspecting their numeral distributions, and rest on (1) assumptions about the probability distribution of results in the absence of fraud or malfeasance, and (2) null hypothesis significance testing. DBEF has been advanced mainly by Professor Walter Mebane of University of Michigan, and his DBEF toolbox is considered to be the standard.

D. Numeral distributions in conventional digit-based election forensics

49. Conventional DBEF methods are based on two general observations. First, many natural processes produce quantities in which the numerals are distributed according to Benford's Law. Hill (1995) proved that in numbers drawn from a random mixture of distributions their numerals converge to a logarithmic distribution. Under Benford's Law, in the decimal number system the distribution of numerals approaches uniformity with digit order, and is practically uniform already for the fourth digit. The second observation is that human subjects of varied backgrounds do not generate numbers that follow Benford's Law when they are asked to generate "random" numbers in their heads (see Nickerson (2002) and Beber and Scacco (2012)). Combining these two observations, we expect numbers generated by accurate elections to conform with Benford's Law, while numbers invented by fraudsters should not. This is the strong distributional assumption of digit-based election forensics.

50. Two lines of criticism arise. The first questions the assumption that numbers in accurate elections should follow Benford's Law. For example, laws and regulations may set limits on the number of voters per polling station (or ward, precinct, etc.). This affects the frequencies of numerals on the first digit of vote counts. For this reason, conventional DBEF methods generally focus on the second digit or higher-order digits.

[1] Consider the hypothesis that the terminal digits are random, independent, and have a 10% chance of being equal to 0, 1, 2, ..., or 9. The chi-square test involves comparing the observed frequency of each digit to its expected frequency if that hypothesis were true--namely, 10% of the number of measurements. The chi-square test sums the squares of the differences between the observed and expected frequencies, each divided by its expected frequency. That sum is an overall measure of how the ten observed frequencies differ from their expected values. If the hypothesis is true, the sum is expected to be relatively small. If, on the assumption that the hypothesis is true, the sum is surprisingly large, that is evidence that the hypothesis is not true.

51. That does not completely solve the problem. As Shikano and Mack (2011) show using German federal elections and simulations, even accurate results can appear far from Benford's Law, for reasons other than irregularities. Furthermore, irregularities where not caused by fraud can produce numbers that do not follow Benford's Law. For example, human vote counters whose mathematical skills are weak may favour certain digits (i.e., "round off") to simplify computations, without intentionally favouring any of the candidates. The second line of criticism targets the assumption that all irregularities—or even all deliberate fraud—will result in violations of Benford's Law. Unlike human subjects in laboratory experiments, some electoral fraudsters know what features of the numbers will be inspected, and also have access to tools to generate numbers that will pass statistical tests based on Benford's law.

52. One alternative to conventional DBEF involves "learning" the distribution of irregularity-free results from the data. This is especially appealing if there is a set of elections available that is believed to be accurate, for elections that took place under conditions similar to those of the election under scrutiny. This allows the strong distributional assumption of conventional DBEF to be relaxed: the irregularity-free numeral distribution does not need to be known ex ante, and does not need to be Benford's Law. Instead, the test relies on weaker assumption that the numeral distributions for accurate elections differ from those of elections with irregularities.

53. There are two strategies for implementing such methods. First, one can treat the elections believed to be accurate as a training set, estimate the features of the reference distribution from them, and compare the evaluated elections to those estimates. In that vein, Cantu and Saiegh 2011 use synthetic data to obtain the features of accurate elections. Alternatively, one can evaluate whether both the accurate elections and the elections under scrutiny can be fit adequately using the same numeral distribution (Medzihorsky 2015). The second option in effect means that the task is no longer to classify whether a set of returns has irregularities or not, but rather whether the distribution of the inspected returns is "close enough" to those believed to be accurate.

E. Null hypothesis testing in conventional digit-based election forensics

54. In conventional DBEF, the observed numeral distribution is compared to a distribution that is hypothesised to have generated the results if there were no irregularities. The comparison takes the form of a statistical significance test of the statistical null hypothesis that the data were drawn from that hypothesized distribution. Null hypothesis significance testing (NHST) is a workhorse of applied statistics and has been fruitfully applied to problems in a very broad variety of domains. Applying NHST generally entails computing a "test statistic" from the data, and comparing its observed value to the distribution of values the test statistic would have if it were applied to many samples of the same size, on the assumption that the null hypothesis is true. This comparison is summarized as a p-value, a measure of how surprising the observed data would be if the null hypothesis were true. (Smaller p-values are stronger evidence that the null hypothesis is false.)

55. The p-value is *not* the probability that the null hypothesis is true, a common misconception. Rather, it is the probability of observing data as "unusual" as the actual data, computed on the *assumption* that the null hypothesis is true. The last step of hypothesis testing is to compare the p-value to a (pre-determined) threshold of statistical significance, such as 5%, 1%, or 0.1%, to decide whether to conclude that the election results reflect irregularities (which occurs if the p-value is sufficiently small). If the null hypothesis is true (i.e., if the election results are accurate) and many repeated samples are taken and inspected, the null hypothesis will be falsely rejected for a known expected fraction of samples.

56. For NHST to be useful, among other requirements, the researcher must select the test statistic and the threshold for the p-value so that they fit the question asked of the test and

the decision to be made based on it. This is especially important because the "power" of tests—the ability of a test to correctly conclude that the null hypothesis is false when a particular alternative hypothesis is true—is limited by the sample size.

57. In particular, DBEF tests can yield different conclusions for two samples in which the numerals appear with the same relative frequencies, but sample sizes differ. The null hypothesis of no fraud is more likely to be rejected for larger sample sizes if the data have the same probability densities over digits and significance threshold.

58. In principle, the significance level should reflect the costs and benefits of correct and incorrect decisions (true and false positives, and true and false negatives). In practice, such informed threshold choices are rare. Instead, practitioners usually pick a conventional significance level, such as 0.05, 0.01, 0.001, or 0.0001. Pre-specifying the level may help ensure that the choice is not engineered to deliver a preferred decision. In the context of election forensics, it is perhaps harder than usual to pick a good level, because elections are typically carried out in intervals of years as one-off events, and much hinges on the outcome and its perceived legitimacy. This is especially true for new or troubled democracies, which are more likely to be inspected for signs of electoral fraud. It is impossible to know in advance what costs and benefits follow from raising flags for accurate elections and for not raising them for elections with irregularities.

F. From "Are the results clean?" to "How much fraud was there?"

59. Broadly speaking, NHST seeks to classify election results as accurate or irregular (although the null hypothesis requires a specific model for how accurate results are generated).

60. In practice, this qualitative question is less interesting than the quantitative one of how much irregularity affected the results, and in particular whether irregularities altered the electoral outcome. Medzihorsky (2015) proposed a latent class to digit based election forensics. In this approach, the observed numerals are sorted into two classes: accurate and irregular. The classes are latent in the sense that their memberships are not observed and have to be inferred from the data.

61. Conventional DBEF is based on the assumption that the exact shape of the numeral distribution of accurate results is presumed known ex ante, but nothing is known about the shape of the numeral distribution for irregular results, beyond the fact that it is different. In other words, while the distribution of accurate results is pre-specified, but the distribution of irregular results is not.

62. Latent-class DBEF decomposes the observed results into the largest possible set that is described well by Benford's Law (or another specific distribution, such as the uniform) and the smallest possible set that is not. In this way, it gives an estimate of the smallest possible fraction of numerical results that do not appear to be accurate.

63. Within this framework, two different interpretations of the residual class are possible that can lead to different numeric values for the same data. The first is that the residual class is composed of numerals that need to be *removed* to obtain results consistent with the presumed distribution of accurate results. The second is that the residual class is composed of numerals that would need to be *changed* in order to obtain a distribution consistent with the assumed distribution of accurate results. The first interpretation leads to a special case of the Rudas et al. (1994) "mixture index of fit." The second interpretation leads to a measure related to the Gini (1914) dissimilarity index.

64. The latent class approach does not require the conventional assumption that the numeral distribution of accurate results is known ex ante, provided there is a set of returns strongly

believed to be accurate, in addition to the results under examination. In that case the latent class approach sorts the digits into a group that describes as many as possible of all the observed numerals, and groups the rest into a residual group.

G. Approaches not based on digits

65. Klimek et al. (2012) propose a statistical approach to electoral irregularity detection that rests on inspecting the joint distributions of turnout and incumbent's (or winner's) shares, rather than the distribution of numerical results. In this approach, the observed joint distribution is inspected for the presence of two components: one containing accurate results and one in which both the turnout and the incumbent's share are considerably higher than elsewhere, which could be caused by malfeasance, such as ballot stuffing, or other irregularities. A related approach has been proposed by Rozenas (2017). His approach rests on inspecting the distribution of incumbent's (or winner's) share to identify whether some reported shares (such as 0.5, 0.6, or 0.75) are anomalously frequent.

VI. Limitations

66. There are two important reservations about forensic statistical procedures when analysing election data for irregularities. However, if election results are available, statistical analyses can identify suspicious patterns, and allow analysis of potential instances of fraud in a more comprehensive way than election observation missions can.

67. Given suitable data, statistical methods can cover the entire territory, and reach down to every single polling station. However, statistical anomalies can occur for many reasons, not only fraud or error, so positive statistical test results can never be considered a direct proof of irregularities, but only evidence complementing other, qualitative assessments, for instance through observation missions. Manipulation of elections or election results or the occurrence of outcome-altering errors can only be proved through direct evidence, such as a rigorous audit against trustworthy, voter-verified paper records.

68. Conversely, statistical procedures can miss gross errors and fraud, and are only sensitive to a limited range of problems in the electoral process. Therefore, we cannot conclude from the absence of statistically remarkable anomalies that elections were well administered and free of interference, manipulation, or outcome-altering errors.

69. There are four major limitations to statistical procedures as tools to detect manipulations in elections:

70. First, many other factors affect election results, and they vary geographically, potentially producing patterns similar to those related to fraud. Concentrations of votes in small regions, in conjunction with high turnout, may result from the territorial concentration of socio-economic or cultural minorities, the prominence and mobilising power of local candidates, or small communities (Coleman, 2004). Even in old democracies, turnout tends to correlate with partisan vote (Grofman et al., 1999). Uneven distributions of digits can result from the design of electoral districts, candidate nomination strategies, or altering voting strategies across electoral constituencies (Cain, 1978; Bochsler, forthcoming; Mebane, 2013). Also, "innocent errors" by election administrators or poll workers may be indistinguishable from fraud by statistical methods.

71. Second, statistical tools for election forensics are always based on probabilistic tests. The patterns they flag as anomalous are unlikely to occur randomly according to their presumed model of accurate elections, but are not impossible. If the same statistical test is applied to a large set of elections that were all accurate (according to the underlying model), we should

expect some of the elections to be flagged as suspicious, because suspicious patterns do arise by chance. Similarly, if many different tests for different types of irregularity are conducted for the same elections, it is likely that one of these tests will show a “false positive” (Leemann and Bochsler, 2014).

72. Third, the electoral process has many steps, e.g., candidate registration, compilation of the register of voters, and ballot counting. Each step can suffer from irregularity (in different forms and shapes). Therefore, statistical tools cannot test for just one form or source of irregularity. While there are sophisticated tests for some instances and forms, tests for others are much more rudimentary, and for many steps in the election process, no systematic data and/or tests are available.

73. Again, statistical tests are always based on probabilistic assumptions about how accurate results are generated, and therefore, irregularities can go undetected by chance, or because they are not pervasive or large enough to be detected. For both reasons, (multiple) statistical tests that do not flag any surprising patterns should never be interpreted as affirmative evidence that elections are accurate.

74. Fourth, Kalinin and Mebane (2017) report patterns in electoral results that raised the suspicion that an electoral commission adapted its manipulations in order commit fraud in a way that standard statistical procedures will not detect. If a standardised package of statistical procedures is established to vet electoral results (e.g. Hicken & Mebane 2015), then fraudsters can take countermeasures to ensure that their work will pass the tests. For instance, by randomising some of the manipulated numbers, statistical tests that are based on numeral distributions will no longer identify these results as irregular.

75. Many socio-geographical or political patterns also lead to vote distributions that resemble election fraud, for instance, producing an association between vote share and turnout. Differences in turnout and in party identification are often due to differences in the political behaviour of social groups. For instance, in the United States, whites tend to be characterized by higher turnout, and stronger support for Republicans. Hence, a positive correlation of the two is not necessarily evidence of error, ballot box stuffing, or other election malpractice (Grofman et al., 1999).

76. Both turnout and partisan vote patterns are influenced by the local popularity of local candidates and their ability to mobilize voters. Studies have found that small communities might have more cohesive preferences, and more uniform turnout pattern, leading to peaks in both partisan votes and turnout (Coleman, 2004). Patterns of strategic behaviour will often vary between different (geographic) electoral districts. For instance, small parties renounce running in districts with a narrow race, or some strategic voters defect from weak candidates (Cain, 1978, Bochsler, forthcoming). This will lead to deviations from Gaussian normal distributions (Mebane, 2013). Statistical tests can take into account demographics of the electoral wards, earlier electoral results, or institutional variance, where data are available. However, other factors can be overlooked, misunderstood, or not identified due to a lack of data at a sufficiently disaggregated level.

VII. Conclusion

77. There are statistical procedures to identify anomalies in election results that might be caused by fraud, manipulation, or gross error. However, the anomalies are not necessarily due to such irregularities, but possibly due to other socio-geographical or political-strategic processes. This is a common problem in the social sciences, which often relies on observational data rather than controlled experiments.

78. There are a number of approaches to take into account potential outside influences, and thereby potentially to improve the validity of statistical procedures used in election forensics—but none is perfect. Hicken and Mebane (2015) provide a package of statistical tests for anomalies in election results. They do not merely apply multiple tests in parallel; they analyse whether several indicators of irregularities agree, arguing that multiple indicators strengthen the evidence. However, particular attention must be paid to the nature of alternative processes that might produce similar anomalies; e.g., variance in the strategic context of an election between districts might affect many tests, and affect both the partisan distribution of the results and the turnout. Furthermore, because some irregularities in the election process are only identifiable by one test, they will not be corroborated by multiple tests. And, as mentioned previously, the null hypotheses tested involve rather stylized, counterfactual assumptions about how accurate election results would be generated. To reject the stylized model is not equivalent to showing that there was an irregularity. It merely means that the data would be surprising if the stylized model were true.

79. Hicken and Mebane (2015) also consider whether anomalies in election results are geographically clustered. This might indicate malpractice in certain parts of the territory (Hicken and Mebane, 2015: 13, Kobak et al., 2016a). Indeed, there is a history of electoral fraud being particularly conducted in the periphery of countries, and/or in regions with different ethnicities. Historically and still today, there are cases where some parts of a country are ruled by authoritarian sub-national governments that do not allow the same political freedoms as other regions. These regions are particularly vulnerable to electoral manipulations, inflated turnout, and a lack of political pluralism (Snyder, 2001). However, factors unrelated to election manipulation can lead to extremely different distributions in the election results in different geographic regions, and to bimodal distributions of the results. For instance, candidates generally win more votes in their home regions, and regions usually differ politically due to a different socio-economic or ethnic composition (Caramani, 2004). Peculiarities in turnout might be due to parallel local/regional elections, local weather conditions, or differences in administrative practices and rules across regions.

80. Leemann and Bochsler (2014) rely on previous information about irregularities in election administration. In the referendum they analyse, a small number of municipal election administrations were found not to have followed the rules, and destroyed the ballots of a controversial referendum before a recount could take place. Relying on previous information about irregularities, statistical tests were run separately for wards where irregularities were known to have occurred and for those where elections were conducted according to the rules. According to the test logic, if the results of the statistical tests conducted on the wards with no irregularities do not show any anomalies, this offers additional confidence that natural processes did not produce misleading evidence of electoral malpractice. The information used to identify wards that are less vulnerable to irregularities might also be based on other pieces of evidence, e.g., evidence collected by election observation missions.

81. Combining information from election observation missions with statistical evidence, Hyde (2011) assesses how the electoral results differ between polling stations that were subject to an international election observation and those that were not observed. Her analysis of elections in Armenia is strengthened by the fact that the polling stations observed were selected randomly. While she finds that polling stations with no election observations reported lower vote shares for the opposition, this might not result from electoral irregularities. In particular, it could be an indirect effect of the presence of the observation mission on election behaviour.

VIII. References

Alvarez, R. M. & Boehmke, F. (2008). Correlates of Fraud: Studying State Election Fraud Allegations. In Alvarez, R. M., Hall, T. E. & Hyde, S. D. (eds.) Election Fraud: Detecting and Deterring Electoral Manipulation. Washington DC, Brookings Institution Press.

Beber,B., Scacco,A. (2012). What the numbers say: A digit-based test for election fraud. Political Analysis, 20 (2), 211-234.

Bochsler, D. (forthcoming). The strategic effect of the plurality vote at the district level. Electoral Studies.

Brancati, D. & Snyder, J. (2013). Time to Kill: The Impact of Election Timing on Post conflict Stability. Journal of Conflict Resolution 57: 822-853.

Cain, B. E. (1978). Strategic Voting in Britain. American Political Science Review 22(3): 639-655.

Cantú, F., & Saiegh, S.M. (2011). Fraudulent Democracy? An Analysis of Argentina's" Infamous Decade" Using Supervised Machine Learning. Political Analysis, 19 (4), 409-433.

Caramani, D. (2004). The Nationalization of Politics. The Formation of National Electorates and Party Systems in Western Europe, Cambridge, Cambridge University Press.

Cesid (2002). Izborni zakon i nacionalne manjine. Beograd, Cesid.

Coleman, S. (2004). The Effect of Social Conformity on Collective Voting Behaviour. Political Analysis, 12: 76-96.

Davis-Roberts, A. & Carroll, D. J. (2014). Assessing Elections. In Norris, P., Frank, R. W. & Martínez I Coma, F. (eds.) Advancing Electoral Integrity. Oxford, Oxford University Press.

Daxecker, U. E. (2012). The cost of exposing cheating: International election monitoring, fraud, and post-election violence in Africa. Journal of Peace Research 49(4): 503-516.

Deckert, J., Myagkov, M. & Ordeshook, P. (2011). Benford's Law and the detection of election fraud. Political Analysis 19(3): 245-268.

European Commission for Democracy Through Law (Venice Commission) (2002). Code of Good Practice in Electoral Matters – Guidelines and Explanatory Report (CDL-AD(2002)023rev2).

European Commission for Democracy Through Law (Venice Commission) (2010). Report on Figure Based Management of Possible Election Fraud. Study No. 583/2010. Strasbourg, Council of Europe (CDL-AD(2010)043).

Freedman, D.A., (2010). Diagnostics Cannot Have Much Power Against General Alternatives, in Statistical Models and Causal Inference, D. Collier, J.S. Sekhon, and P.B. Stark, eds., Cambridge U. Press, NY.

George, J. A. (2014). Can hybrid regimes foster constituencies? Ethnic minorities in Georgian elections, 1992-2012. Electoral Studies 35: 328-345.

Gini, C. (1914). Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazione statistiche. Atti del Reale Instituto Veneto di Scienze, Lettere ed Arti (Series 8) 74:185–213.

Grofman, B., Owen, G. & Collet, C. (1999). Rethinking the partisan effects of higher turnout: So what's the question? Public Choice 99: 357-376.

Hicken, A. & Mebane, W. R. J. (2015). A Guide to Election Forensics. Ann Arbor (MI), University of Michigan.

Hill,T.P. (1995). A statistical derivation of the significant-digit law. Statistical Science, 10 (4),354-363.

Hyde, S. D. (2007). The Observer Effect in International Politics. World Politics 60(1): 37-63.

Hyde, S. D. (2011). Catch Us If You Can: Election Monitoring and International Norm Diffusion. American Journal of Political Science 55(2): 356-369.

Hyde, S. D. & Marinov, N. (2011). Information and Self-Enforcing Democracy: The Role of International Election Observation. New Haven, Yale University.

Kalinin, K. and W.R. Mebane, (2017). When the Russians fake their election results, they may be giving us the statistical finger, The Washington Post, 11 January 2017. https://www.washingtonpost.com/news/monkey-cage/wp/2017/01/11/when-the-russians-fake-their-election-results-they-may-be-giving-us-the-statistical-finger/ (last retrieved 6 May 2017).

Kelley, J. G. & Kolev, K. (2010). Election Quality and International Observation 1975-2004: Two New Datasets. Duke University.
Klimek, P., Yegorov, Y., Hanel, R., & Thurner, S. (2012). Statistical detection of systematic election irregularities. Proceedings of the National Academy of Sciences, 109 (41), 16469-16473.
Kobak, D., Shpilkin, S. & Pshenichnikov, M. S. (2016a). Integer Percentages as Electoral Falsification Fingerprints. The Annals of Applied Statistics 10(1): 54-73.
Kobak, D., Shpilkin, S. & Pshenichnikov, M. S. (2016b). Statistical fingerprints of electoral fraud? Significance 13(4): 20-23.
Kofi Annan Report (2012). Deepening Democracy: A Strategy for Improving the Integrity of Elections Worldwide, published by the Kofi Annan Foundation.
Laakso, L. (2007). Insights into Electoral Violence. In Basedau, M., Erdmann, G. & Mehler, A. (eds.) Votes, Money and Violence. Political Parties and Elections in Sub-Saharan Africa. Scottsville, Nordiska Afrikainstitutet / University of Kwazulu-Natal Press.
Leemann, L. & Bochsler, D. (2014). A Systematic Approach to Study Electoral Fraud. Electoral Studies 35: 33-47.
Lindeman, M., and P.B. Stark, (2012). A Gentle Introduction to Risk-Limiting Audits. IEEE Security and Privacy, 10 , 42–49. Preprint: http://www.stat.berkeley.edu/~stark/Preprints/gentle12.pdf.
Lindeman, M., (2015). Misadventures in Election Forensics: A Note on Choquette & Johnson 2012,
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2689843,
http://dx.doi.org/10.2139/ssrn.2689843 (last retrieved 6 May 2017).
Luechinger, S., Rosinger, M. & Stutzer, A. (2007). The Impact of Postal Voting on Participation: Evidence for Switzerland. Swiss Political Science Review 13(2): 167-202.
Mebane, W. R. J. (2008). Election forensics: the second-digit Benford's law test and recent American presidential elections. IN Alvarez, M., Hall, T. E. & Hyde, S. D. (eds.) Election Fraud: Detecting and Deterring Electoral Manipulation. Washington (DC), Brookings Institution.
Mebane, W. R. J. (2013). Election Forensics: The Meanings of Precinct Vote Counts' Second Digits. Summer Meeting of the Political Methodology Society. University of Virginia.
Medzihorsky, J. (2015). Election Fraud: A Latent Class Framework for Digit-Based Tests. Political Analysis 23(4): 506-517.
Myagkov, M., Ordeshook, P. & Shakin, D. (2009). The Forensics of Election Fraud. Russia and Ukraine, Cambridge, Cambridge University Press.
Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review 109(2).
Norris, P. (2015). Why Elections Fail, Cambridge, Cambridge University Press.
Norris, P., Elklit, J. & Reynolds, A. (2014). Methods and Evidence. In Norris, P., Frank, R. W. & Martínez I Coma, F. (eds.) Advancing Electoral Integrity. Oxford, Oxford University Press.
Rozenas, A. (2017). Detecting Election Fraud from Irregularities in Vote-Share Distributions. Political Analysis , 25 (1), 41-56.
Rudas, T., C. Clogg, and B. Lindsay. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. Journal of the Royal Statistical Society. Series B (Methodological) 56(4):623–39.
Schedler, A. (2002). The Menu of Manipulation. Journal of Democracy 13(2): 36-50.
Shikano, S., & Mack, V. (2011). When Does the Second-Digit Benford's Law-Test Signal an Election Fraud?. Jahrbücher für Nationalökonomie und Statistik , 231 (5-6), 719-732.
Stark, P.B., and D.A. Wagner, 2012. Evidence-Based Elections. I EEE Security and Privacy, 10, 33–41. Preprint: http://www.stat.berkeley.edu/~stark/Preprints/evidenceVote12.pdf
Snyder, R. (2001). Scaling Down: The Subnational Comparative Method. Studies in Comparative International Development 36(1): 93- 110.
Wand, J. N., Shotts, K.W., Sekhon, J., S., Mebane, W. R. J., Herron, M. C. & Brady, H. E. (2001). The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida. American Political Science Review 95(4): 793-810.